Introduction

The World Health Organization (WHO) reports that 80% of deaths related to cardiovascular diseases (CVD) are attributed to heart attacks. Early detection of individuals who are at high risk of developing CVD can help reduce mortality rates. In this project, I aim to determine the proportion of patients who have a high likelihood of developing CVD using LinearDiscriminantAnalysis() and machine learning techniques to classify the records in a dataset based on the values of their input fields.

Data

I used the Framingham Heart Study dataset, which contains information on 4,434 patients from the Framingham Heart Study. The dataset includes demographic information, medical history, and laboratory test results. I performed data cleaning by removing missing values and scaling the data to ensure that all features are on the same scale.

Methods

I chose LinearDiscriminantAnalysis() because it is a classification algorithm that works well for high-dimensional datasets like the Framingham Heart Study dataset. It projects the data onto a lower-dimensional space while maximizing the separation between classes. I split the data into training and testing sets.

Results

My model achieved an accuracy of 92.11%, precision of 100%, recall of 40%, and an F1 score of 57.14% on the test set. The confusion matrix shows that the model correctly classified 33 individuals as high risk for CVD and misclassified 4 individuals as low risk. The ROC curve shows that the model has a good trade-off between true positive rate and false positive rate.

Discussion

My model performed well in predicting high-risk individuals for CVD. However, the recall rate was lower than expected, indicating that some high-risk individuals were misclassified as low risk. Possible improvements include feature engineering, using different machine learning techniques, and adding more data to the dataset.

Conclusion

In conclusion, our model can identify individuals who are at high risk for developing CVD with an accuracy of 92.11%. Early detection of high-risk individuals can lead to timely intervention and prevention of CVD-related deaths. Further research can improve the performance of the model and ultimately contribute to better public health outcomes.

References

Framingham Heart Study dataset. National Heart, Lung, and Blood Institute. Retrieved from https://biolincc.nhlbi.nih.gov/studies/framcohort/